Too Many Questions? What Can We Do? : Multiple Question Span Detection

Prathyusha Danda
LTRC, KCIS
IIIT-Hyderabad
danda.prathyusha@research.iiit.ac.in

Brij Mohan Lal Srivastava
INRIA
France
brij.srivastava@inria.fr

Manish Shrivastava
LTRC, KCIS
IIIT-Hyderabad
m.shrivastava@iiit.ac.in

Abstract

When a human interacts with an information retrieval chat bot, he/she can ask multiple questions at the same time. Current question answering systems can't handle this scenario effectively. In this paper we propose an approach to identify question spans in a given utterance, by posing this as a sequence labeling problem. The model is trained and evaluated over 4 different freely available datasets. To get a comprehensive coverage of the compound question scenarios, we also synthesize a dataset based on the natural question combination patterns. We exhibit improvement in the performance of the DrQA system when it encounters compound questions which suggests that this approach is vital for real-time human-chatbot interaction.

1 Introduction

Traditional question answering systems retrieve information from a knowledge-base in accordance with what is being asked in a user utterance. Questions in these systems are queried in a single question format, such that there is only one question per utterance. However, most of these systems suffer in question-answering accuracy, especially when speakers embed multiple questions within the same utterance. QA systems like DrQA by (Chen et al., 2017) do not perform well in cases when the user utterance contains more than one question. The performance of such systems is generally suboptimal, because the answers are generated through the assumption that exactly one question is embedded within one complete utterance. In other words, the entire utterance is processed as a single question. We propose a front end for question answering systems that detects question spans within the utterance, especially when multiple questions are compounded together by the user. We report accuracies comparable within the utterance.

In order to establish the need for such a front end, we conduct a preliminary study by first retrieving all the question instances in the Ubuntu dialogue corpus. One such instance from Ubuntu dialogue corpus is: *why would you recommened arch-linux ? how is it comparable to debian or ubuntu ?*. The utterance might contain more than one question based on the number of contiguous question mark clusters. Such questions exhibit compound question scenario. These questions are usually asked to avoid setting up the context again or for brevity in the dialog. We encountered several patterns for compounding the questions. In order to obtain compound questions, we artificially synthesized the single question instances into relevant compound questions with the most frequent question combination patterns seen earlier. We call our dataset CompoundQA. We evaluated our Multiple Question Span Detection (MQSD) model by using it as the pre-processor to the DrQA system. We observe increase in performance of the system over the compound questions data.

The rest of the paper is organized as follows: Section 2 surveys the related work, Section 3 gives the available datasets description. Section 4 details our approach of creating Compound QA dataset and model description. Question prediction analysis is

Corpus	Total samples	Avg. Sent. Length	Median
Ubuntu	273,133	10	8
SQUAD	98,424	11	11
WikiMovies	107,640	8	8
WebQuestions	5,817	8	8

Table 1: Data statistics after pre-processing

done in Section 5. Section 6 presents the evaluation along with results and Section 7 concludes the papers with remarks on future work.

2 Related Work

Understanding each part of the text written or spoken by the user is essential to QA systems. Once such an understanding is established, relevant information can be easily retrieved. There have been several attempts ((Zhang and Lee, 2003),(Stolcke et al., 2000)) to classify written text into several semantic tags (such as dialog acts, rational speech acts, etc.) for a better response. We specifically deal with questions embedded within Ubuntu chat logs. Although there has not been an attempt to discover several questions compounded together in a single utterance, there have been two such works to identify questions within tweets. Li et al. (2011) claim theirs to be the first such work and they employ rule-based as well as support vector machines to classify tweets containing questions. Dent and Paul (2011) proposed another technique based on comprehensive linguistic parsing of tweets and then classifying them as questions. In the study conducted by (Wang and Chua, 2010) to mine syntactic and sequential patterns within community QA data to classify questions in Yahoo! Answers dataset. These described techniques do not detect question boundary but, only classify a text as question or not.

3 Data

We use four datasets, one of which is a dialog corpus and the remaining are open domain QA datasets. Ubuntu dialogue corpus is used to understand the patterns of asking multiple questions within a single utterance when in conversation with another human. We build an artificial corpus using open domain QA datasets - SQUAD, Wiki Movies and Web Questions based on these observations.

3.1 Ubuntu Dialogue Corpus

The Ubuntu Dialog Corpus (Lowe et al., 2015) is an archive of two-person conversations extracted from the Ubuntu chat log. It contains around 1 million multi-turn dialogues, which consists over 7 million utterances, composing 100 million words. We extract only those utterances which contain question marks ('?'). We assume that question spans occur in all of these extracted utterances. Table 1 gives the total number of extracted utterances, which will be used as training data for our experiments.

Here are a few instances of questions found in Ubuntu dialogue corpus.

- *how to acces a file with a path if i get permission denied ???*
- *you mean the dpkg-reconfigure command ? where is it stuck at ? if it is indeed stuck*
- *has anybody tried connecting your phone and PC via bluetooth ? Did you get it working ?*

3.2 Open domain QA datasets

We use three open domain QA datasets, namely SQuAD, WikiMovies and WebQuestion to build our artificial compound question corpus.

The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) is a reading comprehension dataset. It comprises of over 100,000 questions based on Wikipedia articles, the corresponding answer is a segment of text from the related relevant passage.

(Berant et al., 2013) developed the WebQuestion dataset to answer questions from the Freebase knowledge base, by crawling questions using Google Suggest API. The answers for these questions were then obtained using Amazon Mechanical Turk.

WikiMovies (Miller et al., 2016) originally created from OMDb and MovieLens databases contains 96k question-answer pairs in the movie domain.

Following are few question samples from the above datasets.

- *Which prize did Frederick Buechner create?*
- *who did the philippines gain independence from?*

- *What movies can be described with chris noonan?*

4 Approach

Our approach comprises of understanding the natural question combinations that occur in the Ubuntu dialogue corpus and build a model to identify the question spans in an utterance. As there presently exists no such compound question dataset, we create a dataset CompoundQA which consists of compound questions, and train and test our model on it.

4.1 CompoundQA dataset creation

Ubuntu dialogue corpus consists of utterances which have only one question span in them or more than one question spans (Section 3.1). We observe that most of the utterances have more than one question in them. An interesting observation is that the number of utterances with two question spans is more frequent as compared to multiple question spans instances. This shows the general human behavior of asking two questions is common in a natural conversation scheme.

This shows that in real life scenarios compound questions are created by using discourse connectives. We also observe the propensity of dropping this conjunctions. As a simplistic strategy, we combine two question spans randomly chosen from the existing open domain QA datasets by connecting them with discourse connectives such as 'and', 'also' or sometimes simply the '?' acting as a connective. The mentioned conjunctions are used with uniform probability to generate the data. Naturally this strategy does not take semantic similarity or semantic content into account. Also this does not make any changes to the syntactic structure of the question spans apart from adding the discourse connectives. In Section 6, we show the improvement in performance of the DrQA system on training the model using CompoundQA dataset.

We take all the utterances which have '?' in them to create Ubuntu With Question Mark (UWQM) dataset. To capture the question span in the utterances we created labels for extracted and preprocessed Ubuntu dialogue corpus samples (Section 3) using the standard BIO format. The start of the question span is tagged with 'B-Q' and all the following tokens which are part of the question are tagged as 'I-Q' and the non-question tokens are tagged 'O'.

The following are few examples of tagged ubuntu data:

- **Question:** you mean the dpkg-reconfigure command ? where is it stuck at ? if it is indeed stuck
  Tag: B-Q I-Q I-Q I-Q I-Q I-Q B-Q I-Q I-Q I-Q I-Q I-Q O O O O O
- **Question:** how to acces a file with a path if i get permission denied ? ? ?
  Tag: B-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q O O

To emulate the user behavior of dropping '?', we replace all the question marks in the extracted utterances with '.' to create Ubuntu Without Question Mark (UWoQM) data. We label this no question mark data using BIO format.

We take 20000 samples each from SQUAD and Wiki Movies dataset, and 5000 samples samples from WebQuestions, to construct the CompoundQA dataset. From these 25000 samples, 3000 samples are randomly picked, and another 3000 samples are picked and '?' is dropped. This sampling was done without replacement. In addition to these, the compound questions are created by combining any two randomly picked questions with 'and', 'also' or none.

Four patterns are followed when creating the compound questions:

1. both the question spans have '?' in them
2. none of the question spans have '?'
3. first question span has '?' followed by a question phrase with no '?'
4. second question span contains a '?' where as the first does not.

From each of the above 4 categories 3000 questions are sampled. All these patterns where constructed taking into account the various possible occurrences. We also introduce noise by tagging some of the utterances incorrectly. Below are few samples from CompoundQA.

- **Question:** Who won the 2011 election
  Tag: B-Q I-Q I-Q I-Q I-Q

Data Id	Data	#train	#dev	#test
1	Ubuntu data with Question Marks	16626	4157	5000
2	Ubuntu data without Question Marks	16626	4157	5000
3	Ubuntu data with and without Question Marks	16337	4085	5000
4	CompoundQA Data	11499	2875	3593
5	CompoundQA data and Ubuntu data without Question Marks data	28126	7033	8594
6	CompoundQA data and Ubuntu data without Question Marks data, with Noise	28126	7033	8594

Table 2: Statistics of training, development and testing data

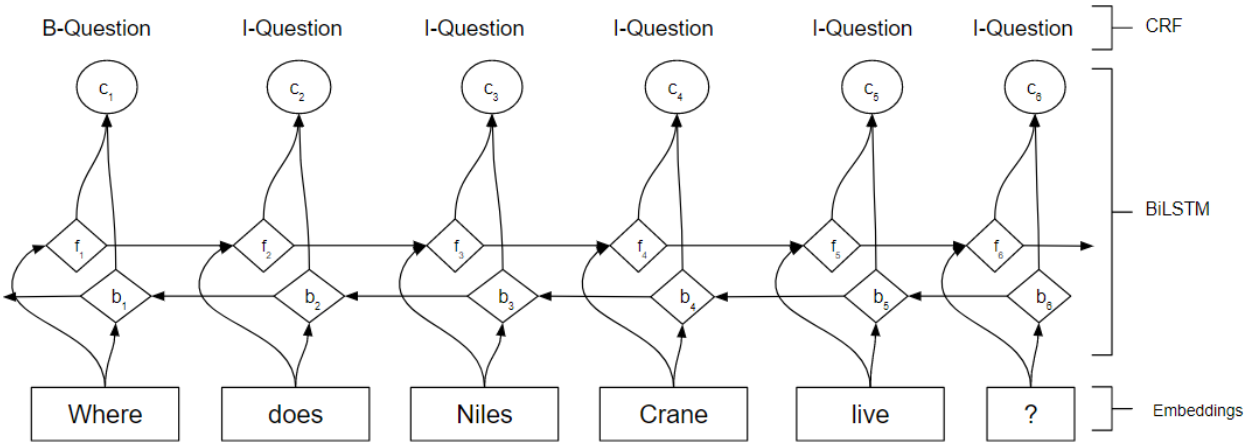

Figure 1: Bidirectional LSTM - CRF architecture for question span prediction.

- **Question:** What decade did herbicides become common ? and how many are believed to have been uprooted by this unrest ?
  Tag: B-Q I-Q I-Q I-Q I-Q I-Q I-Q O B-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q
- **Question:** What is professional wrestling ?
  Tag: B-Q I-Q I-Q I-Q I-Q
- **Question:** On what day did airborne radar help intercept and destroy enemy aircraft for the first time and what will IBM use to analyze weather and make predictions ?
  Tag: B-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q O B-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q I-Q

Table 2 gives the statistics of train, dev and test sets for datasets.

4.2 Multiple Question Span Detection Model

Our sequence prediction model is based on the Bidirectional LSTM-CRF model proposed by (Huang et al., 2015).

The Bidirectional LSTM (BiLSTM) (Graves and Schmidhuber, 2005) is capable of capturing the forward and backward dependencies in a sentence and Conditional Random Field (CRF) (Lafferty et al., 2001) models the whole sentence to generate question span prediction tags. The word embeddings are generated using the procedure explained in (Lample et al., 2016). As per their algorithm, we concatenate the last states of forward and backward pass of a character-level Bidirectional LSTM network trained over the vocabulary. This vector is further concatenated to a pre-trained GloVe (Pennington et

Experiment	Training Data	Testing Data	F1-Score
Experiment-1	Ubuntu data with Question Marks	Ubuntu data with Question Marks	99.44
Experiment-2	Ubuntu data without Question Marks	Ubuntu data without Question Marks	97.50
Experiment-3	Ubuntu data with and without Question Marks	Ubuntu data with Question Marks	99.6
		Ubuntu data without Question Marks	97.49
		Ubuntu data with and without Question Marks	98.5

Table 3: Experiment details with F1-scores on Ubuntu dialogue corpus

al., 2014) word embeddings . The final embedding is provided to the model presented in Figure 1 for question span prediction. In Figure 1 f_i and b_i represent the forward and backward pass states in the sequence. c_i is the context vector used as input to CRF to generate distribution over question BIO tags.

We train and test our model on the Ubuntu dialogue data with '?' in each utterance and observe that the model predicts the question spans with very less error. As in a general scenario the user might drop the '?', we also test the model trained on with '?' data on data without '?' and data which consists of both the cases: with and without '?'

4.3 Experimental Setup

The BiLSTM-CRF architecture is implemented in tensorflow. Pre-trained Common Crawl word embeddings[1] of size 100 were used to initialize the model. Using the training, development and test datasets we construct a vocabularies of words, tags and all the characters present in the data. We load only the vectors of words which are present in our vocabulary to optimize memory usage. The dimension for character embeddings that we trained, is set to 50. We used Adam optimizer (Kingma and Ba, 2014) and dropout (Srivastava et al., 2014) was set to 0.5. The learning rate was set to 0.001 and learning rate decay to 0.9. Hidden embedding dimensions for character and word BiLSTM was set to 50 and 100 respectively. This makes the final word embedding size to be 200-dimensional vector. Batch size of 20 was taken and number of epochs was limited to 30, with an option of terminating if no significant decrease in loss is observed for the three previous epochs.

With the above model parameters, we ran several experiments on different train and test datasets. Individual F1-scores for each dataset are given in Table 3. Experiments 1, 2 and 3 are run on different settings of Ubuntu dialogue data and tested on the corresponding setting. In Table 4, Experiment-4 was trained and tested on CompoundQA dataset. Experiment-5 was trained on Ubuntu data, where question marks were replaced, augmented with the CompoundQA dataset and tested on CompoundQA and Ubuntu dialogue corpus separately. Experiment-6 is similar to Experiment-5 but, noise is introduced in the CompoundQA dataset. We test the model on both CompoundQA and Ubuntu dialogue corpora independently.

5 Question Prediction Analysis

We observe from Experiment-1 that when the model is trained on the Ubuntu data which has question marks at the end of each question span the F1-score is very high. This is because '?' acts as a demarcation for the end of question span and hence the model learns the question spans with more accuracy. To observe the model performance on data without '?' we performed Experiment-2, where the model was trained on data in which question marks were replaced with '.'. The F1-score is less compared to Experiment-1 as the model has to distinguish between the '.' which occurs at the end of question span and all the other occurrences of '.' that might occur anywhere in the sentence. In Experiment-3 the training data is combination of data with and without question marks, it was tested on three datasets. The model does not show increase in the test data with-

[1] https://nlp.stanford.edu/projects/glove/

Experiment	Training Data	Testing Data	F1-Score
Experiment-4	CompoundQA data	CompoundQA data	98.99
Experiment-5	CompoundQA data and Ubuntu	CompoundQA data	99.03
Experiment-6	CompoundQA data and Ubuntu data without Question Marks data, with Noise	CompoundQA data	99.25

Table 4: Experiment details with F1-scores on CompoundQA and Ubuntu dialogue corpus

out '?', but there is an increase in the test data which has '?' as the model was trained on more training data compared to Experiment-1.

In Experiment-4, Table 4, we train and test our model on the CompoundQA dataset. The error cases consisted of question spans with abbreviations or names in them. We observe that the sequence is incorrectly labeled in cases where there is no '?'. To reduce error in these cases we combine CompoundQA with Ubuntu without question mark data and observe an increase in the F1-score as compared to the Experiment-4 when tested on CompoundQA. This increase suggests that the model learns from the natural question spans of Ubuntu data. Experiment-6 results on both the test datasets suggests that inclusion of noise in the training data does not affect the performance of the model.

6 Evaluation and Results

To evaluate our multiple question span detection model, we apply it over an existing question answering system and analyze the performance of the QA system. Recently published work on open domain QA system DrQA, has shown comparative results on various datasets by relying on a unique knowledge resource - Wikipedia. We test our model's performance by applying it over DrQA system.

The existing 4998 samples of WebQuestion dataset (3.2) are used to create 2499 compound questions following the rules listed in Section 4. Each of these 2499 compound questions contain two different question spans. The 2499 compound questions built from the 4998 question samples are stored along with the corresponding 2499 **DrQA predicted answer pairs**.

The predictions of the 4998 single span questions when given to the DrQA system are considered as DrQA predicted answers. In Figure 4 we compare

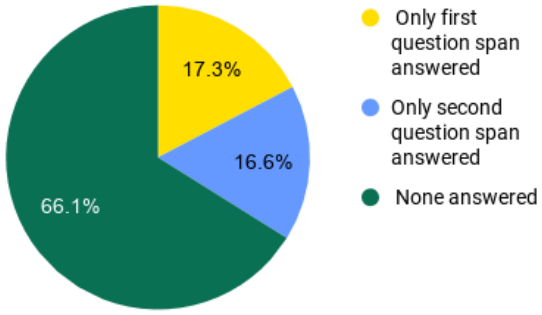

Figure 2: Statistics over DrQA Model.

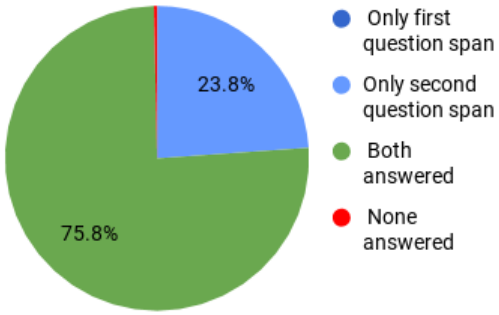

Figure 3: Statistics over MQSD+DrQA Model.

the DrQA predicted answers with the actual human annotated WebQuestion answers, and observe that only 711 questions out of the 4998 questions are answered correctly. *For our analysis we compare our predictions with the DrQA predicted answers.* This relative comparison is done to exclude DrQA model error when calculating MQSD system performance.

Compound questions are given to the DrQA system as input and the obtained predictions are compared with the *DrQA predicted answer pairs*. We observe that for no sample both the answers are predicted correctly. For a few samples either the first question span is answered correctly or the second one. On further analysis we observed that in 433

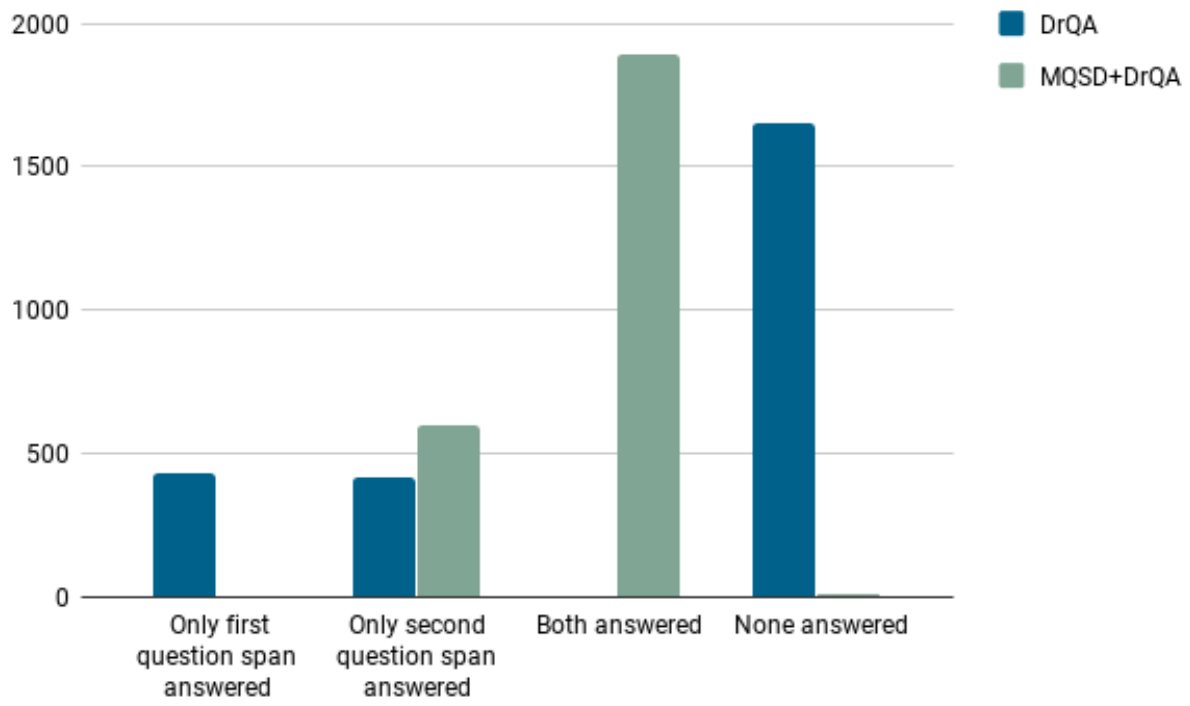

Figure 4: Evaluation details on predicting answers with and without MQSD on CompoundQA dataset

questions the first question span was answered correctly, where as in 413 questions the second question span's answer was predicted and in for no sample both the question spans were answered as shown in Figure 2. 'Only first question span answered' considers all the samples in which the first question span is answered and not the second, same intent applies to the category 'only second question span answered'. By 'Both answered' we take all cases where both the question spans are answered and 'none answered' is where neither the first nor the second question spans are answered.

The first example listed below shows the case where the first question span is answered whereas in the second example the second question span's answer is predicted. In the third example the prediction contains answers for neither of the posed questions. The ground truth is the compound answer predicted by the DrQA system when it is given the two questions in the pair, separately.

- **Question:** what films has scarlett johansson been in also what is monta ellis career high points
  DrQA predicted answer pair: Maggie the Cat, Team leader
  Predicted answer: Maggie the Cat
- **Question:** who is the 2011 heisman trophy winner and what sea does the yangtze river empty into
  DrQA predicted answer pair: Chris Weinke, East China Sea
  Predicted answer: East China Sea
- **Question:** what kind of money does chile use and what percent of mississippi is black
  DrQA predicted answer pair: Chilean peso, 33 %
  Predicted answer: 6.63 %

We perform experiment 6 (Table 4) on compound questions prior to predicting the answers using DrQA. After identifying the question spans in the sample, each question span is separately given to the DrQA system to get the corresponding predictions. We observe that out of the 2499 compound questions, 1894 samples have correct prediction for

both the answers in the pair.

Below are the examples where only the first and second span are answered correctly. In the third example none of the predictions are correct. The "Actual question span" is the expected question spans separated by $ and "Predicted question span" field gives the spans predicted by the MQSD model. The errors observed fall under the cases mentioned in Section 5.

- **Question:** where does niles crane live ? and where did c.s.lewis go to college ?
  Actual question span: where does niles crane live ? $ where did c.s.lewis go to college ?
  Predicted question span: where does niles crane live ? and $ lewis go to college ?
  DrQA predicted answer pair: Manchester, City of Marion
  Predicted answer: Manchester, Punch-Drunk Love"
- **Question:** who speaks farsi and who voiced meg in the pilot ?
  Actual question span: who speaks farsi $ who voiced meg in the pilot ?
  Predicted question span: who speaks farsi and $ who voiced meg in the pilot
  DrQA predicted answer pair: Jeff Jarrett, Mila Kunis
  Predicted answer: Iraj Ghaderi, Mila Kunis
- **Question:** where is located cornell university also when was george h.w . bush elected president ?
  Actual question span: where is located cornell university $ when was george h.w . bush elected president ?
  Predicted question span: where is located cornell university also $ bush elected president
  DrQA predicted answer pair: Manhattan, 1836
  Predicted answer: Ithaca, Martin Van Buren

Figure 2 and Figure 3 summarize the statistics with and without the MQSD model over DrQA. Figure 4 compares with and without MQSD model over DrQA. This summary helps us visualize and compare the nature of error made by the baseline and MQSD system along with the distribution of samples in those error categories.

7 Conclusion and Future Work

We addressed the need for identifying question spans in a user utterance when interacting with a QA system through the analysis of Ubuntu dialogue corpus utterances. Multiple question span detection is posed as a sequence labeling task which we modeled using a Bidirectional LSTM - conditional random field network. We built a simulated compound question dataset CompoundQA using existing open domain QA datasets. The MQSD model was trained and tested on both Ubuntu dialogue utterances as well as CompoundQA dataset. We demonstrate that the present QA systems do not handle multiple question spans and using the MQSD model as a front-end to open domain QA system DrQA boosts it's performance when compound questions are given.

Question span detection is crucial for open domain dialog systems as well. In the open domain dialog systems a user either chit-chats with the system or has a fixed goal. Identifying the question span in goal oriented cases will help the system know the intent of the user and thus help in retrieving relevant information. As a future work, we plan to capture the questions by considering the conversational context as a parameter to MQSD.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Kyle D Dent and Sharoda A Paul. 2011. Through the twitter glass: Detecting questions in micro-text. In *Analyzing Microtext*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Diederik Kingma and Jimmy Ba. 2014. Adam a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Baichuan Li, Xiance Si, Michael R Lyu, Irwin King, and Edward Y Chang. 2011. Question identification on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2477–2480. ACM.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Kai Wang and Tat-Seng Chua. 2010. Exploiting salient patterns for question detection and question retrieval in community-based question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1155–1163. Association for Computational Linguistics.

Dell Zhang and Wee Sun Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 26–32. ACM.